A time series machine learning approach for vigilance classification by non-invasive infrared thermography

Maureen Bosch

5630649

January 10, 2021

Supervisor and first examiner: Dr. Nico Romeijn

Second examiner: Dr. Baptist Liefooghe

Master Artificial Intelligence,

Utrecht University

Abstract

Improvement of vigilance measurement and data analysis is needed to make vigilance classification more applicable in real-world situations. This study aims to evaluate whether non-invasive continuous infrared thermography can be used as a vigilance measurement, whether machine learning models can improve data analysis and whether adding Time Series Analysis can improve vigilance classification. A 10-minute psychomotor vigilance task was conducted with 29 participants. The baseline Generalized Linear Model was compared to a Support Vector Machine model with Radial Basis Function and a Long-Term Short Memory neural network model. Three distal-to-proximal temperature gradients measured by iButtons and an infrared camera were used as predictors to classify vigilance. The Hanley and McNeil test showed no difference between the models and different model predictors. All models classified vigilance around chance level. Future research should include a more distributed participant group and more statistical features.

Keywords: Infrared Thermography, Machine Learning, Thermoregulation, Time Series Analysis, Vigilance Classification

Introduction

Vigilance is a vital physiological signal and is usually defined as *'the ability to sustain attention to a task for a period of time'* (Oken et al., 2006; Parasuraman et al., 1998). Any loss of vigilance during vigilance demanding tasks can have dramatic consequences, such as accidents. Around 10-20 percent of the fatal accidents on the road are ascribed to drivers with a diminished vigilance level caused by fatigue (Aarts et al., 2016; Bergasa et al., 2006). Not only driving, but various occupations with safety-related operations such as air traffic control, rail services, and medical services require continuous sustained high vigilance (Donald, 2008; Masoudian &

Razavi, 2019; Zheng & Lu, 2017). Investigations in industrial operations and transportation have shown that diminished vigilance contributed to serious incidents and accidents (Dinges, 1995). Therefore, detection and prediction of real time vigilance levels is important to provide a safe working environment.

It is difficult to accurately estimate vigilance during vigilance demanding processes. The main reason is that vigilance states are intrinsic mental states that involve temporal evolution rather than a time point (Zheng & Lu, 2017). Furthermore, it is difficult to evaluate mental states without using an intrusive stimulus or behaviour probe. In order to improve vigilance assessment and make detection and prediction more applicable in real-world situations it requires continuous non-invasive measurement with high temporal resolution and a machine learning model data analysis (Davidson et al., 2007; Zhang et al.,2016a; Zheng & Lu, 2017).

Multiple physiological parameters have been indicated to assess vigilance, of which electro-encephalography (EEG) is the most commonly used measurement (Dinges, 1995; Oken et al., 2006; Oudonesom, 2001). A detailed overview of EEG and other physiological measurements to capture vigilance can be found in Oudonesom (2001). The key metrics and their strengths and weaknesses are summarized below.

EEG is the golden standard for vigilance detection because it can directly reflect human brain activity (Zheng & Lu, 2017). It has a high temporal resolution and low-cost non-invasive properties (Zhang et al., 2016b). However, when making EEG user applicable and feasible in, e.g. cars, problems arise. Most data collection requires skin preparation and conductive gel application to secure the sensors to the human skin. This is time consuming, uncomfortable and can be painful for the driver (Lin et al., 2014; Oudonesom, 2001). Wireless and wearable EEG systems reduce these problems by using flexible dry electrodes (Zhang et al., 2016b). However,

it should still be properly attached to the skin and therefore true unobtrusiveness remains a challenge.

There are other continuous physiological parameters that have been shown to correlate with vigilance such as thermoregulation. The thermoregulation system alters skin blood flow which changes heat loss through convection and radiation of the skin (Ghahramani et al., 2016; Kräuchi & Wirz-Justice, 1994; Hasselberg et al., 2013). Heat loss mainly takes place in distal skin locations, such as the hands and feet, because of the high density of arteriovenous anastomoses (AVA) (Charkoudian, 2003; Walløe, 2016). Contrarily, proximal skin locations, relative to the core of the body, have less or no AVA's. For a detailed overview of AVAs and their role in thermoregulation see Walløe (2016).

Multiple studies have shown that spontaneous induced fluctuation in skin temperature support an association with vigilance (Kräuchi et al., 1999; Romeijn et al., 2012). According to Kräuchi et al. (2000) the best predictor to measure vigilance is the gradient between distal to proximal skin temperature (DPG). However, these studies used sensors to measure the skin temperature of the participants, which are invasive.

Infrared thermography (IRT), on the other hand, operates in the infrared band in the electromagnetic spectrum (Abdelrahman et al., 2017). It records the radiating energy that is released from the body which is directly related to skin temperature (Fernández-Cuevas et al., 2015; Ioannou et al., 2014). The advantage of IRT is that these recordings are contact-free. It ensures the isolation of unsystematic data variation such as the user's bias due to their awareness of being monitored, the movement of the sensors or the stressful attachments of sensors to the user's body (Abdelrahman et al., 2017). IRT is therefore non-invasive and more user-friendly than EEG and other thermoregulation measurements.

Thermal imaging has been linked to related cognitive and affective processes such as cognitive (work)load (Abdelrahman et al., 2017; Reyes et al., 2009; Stemberger et al., 2010), driving performance (Reyes et al., 2009), fatigue (Aryal et al., 2017) and stress (Puri et al., 2005), but it is currently unknown if IRT is able to predict vigilance.

Apart from non-invasive data acquisition, further improvement in vigilance assessment could possibly be made in its analysis. Currently, research of thermography on vigilance assessment has focused on finding a trend in thermography on vigilance assessment by using linear regression (Fronczek et al., 2008; Raymann & van Someren, 2007; Romeijn et al., 2012). Their goal was to find a relationship between thermography and vigilance rather than to detect and predict real-time vigilance estimation. This study, however, wants to find a workable and reproducible model to monitor, and possibly warn for, a drop in vigilance. The model needs to automatically detect meaningful patterns in the data, which can be achieved by machine learning (Shalev-Schwarts & Ben-David, 2014). Machine learning is one area of artificial intelligence that attempts to emulate human behaviour. It involves building a model that applies learning methods to find correlations in the data. It can either tackle classification problems or regression problems (Zhao et al., 2008).

This research focuses on a practical application of probable loss detection in vigilance, which is more of a classification problem than a regression analyses. Although linear regression equations can be used for classification (Neapolitan & Jiang, 2018), it cannot predict values perfectly (Gravetter & Wallnau, 2013). Moreover, nonlinearities are often crucial in finding aspects of physiological functions (Marmarelis, 1997), such as skin temperature (Kitney, 1975). Therefore a linear function may be too limiting to capture the relationship between thermoregulation and vigilance.

Support vector machines (SVM) can overcome this problem by using the kernel trick. SVM is a learning model that creates a linear separating hyperplane, but can embed the data

into a higher-dimensional space. Data that is not linearly separable in the original input space can then be easily separable in the higher dimensional space. It creates a non-linear original space and can be seen as an expansion to the linear regression (Russel & Norvig, 2016, pp. 744-748 ; Shalev-Schwartz & Ben-David, 2014, pp 167-177).

Some researchers have used SVM algorithms to predict vigilance (Armanfard et al., 2016; Lin et al., 2014; Wei & Lu, 2012; Zhang et al., 2016a). Armanfard et al., (2016) conducted a vigilance classification experiment by a psychomotor vigilance task (PVT). Using EEG signals extracted from a brain-sensing Muse headband and a SVM with linear kernel function to identify vigilance lapses. It achieved a classification accuracy of 94.6%. These studies used EEG signals as input features to predict and detect vigilance levels. However, it is currently unknown whether machine learning techniques can be used with thermography to classify vigilance.

Lastly, further improvement of vigilance assessment can be made by incorporating the temporal structure of the data. According to Zhang et al. (2016a) the above named studies ignore the time dependency of the vigilance changing process. The studies treat the subject's mental states as independent points and discard the temporal dependency information. The use of independent points tends to be too variable to observe rapid temporary lapses (Davidson et al., 2007). Therefore continuous task performance and psychophysiological measures over time should be analysed instead (van Orden et al., 2000).

This temporal detection can be tackled by time series classification, where the goal is to find a function of an object and time which can detect scenarios of a given class from past and present input values only (Geurts, 2001). Several types of time series classification algorithms have been developed of which Recurrent Neural Networks (RNN) is one of them.

RNN is a kind of artificial neural network which is motivated to compute parallel to the human brain (Hassoun et al., 1995). It is based on a collection of connected units called artificial neurons (Schnupp et al., 2010). In RNN, these connections between units form a cycle which makes it suitable to process continuous data sequences (Sano et al., 2018; Zhang et al., 2016a). RNNs are trained using stochastic gradient descent, which calculates the prediction error and uses this error gradient to update the network weights. A limitation to RNN models is that the model slows down learning or even stops when the gradient signal becomes too small (Sano et al., 2018).

In order to control this gradient, Long Short Time Memory (LSTM) neural networks, a type of RNN, can be incorporated (Hanin, 2018). LSTM is capable of learning long-term dependencies and does not employ fixed memory representations (Davidson et al., 2007). It contains hidden layers with memory cells and an input, output and forget gate. For each input, the function determines to remember or forget the value and when to output the value (Sano et al., 2018; Zhang et al., 2016a). As a result, it can store states over long periods of time and obtain temporal sequence tasks such as machine translation (Sutskever et al., 2014) and speech recognition (Sundermeyer et al., 2012).

Zhang et al. (2016a) incorporated a LSTM model as a temporal encoder to estimate vigilance by EEG and forehead Electrooculogram (EOG). They found that incorporating LSTM models in the data analysis can improve vigilance estimation over SVM models. It is currently unknown if time series classification can be applied for thermography.

Since practical vigilance assessment is of crucial importance to monitor, and possibly warn for, a drop in vigilance, the current study will assess the practical application on non-invasive continuous measurement to detect and predict vigilance levels. The main objectives are to assess if thermoregulatory changes measured by IRT could be used as a continuous non-

invasive measurement, if machine learning models can improve data analysis and whether adding time series can improve vigilance classification.

Methods

Experimental setup

This study used a dataset of skin temperature responses collected during a sustained attention reaction time task. This data was processed to train and test multiple machine learning algorithms to classify vigilance. All procedures complied with the guidelines set out in the Declaration of Helsinki.

Participants

Twenty-nine voluntary participants (four male, Median 21, range: 19-29 years) were recruited within the University of Utrecht and the observers' social network. All participants were informed about the experiment of the study beforehand and provided written informed consent. None of them had any known history of concentration-related disorders and none of them smoked. Since caffeine has shown to affect vigilance (Gilbert et al., 2000) and skin temperature (Quinlan et al., 2000), caffeine intake was prohibited from 3 hours prior to the experiment onward.

Experimental procedure

The procedure took place between 13:00 and 15:00 o'clock, for both skin temperature and vigilance tend to fluctuate most during this time of day (van Marken Lichtenbelt et al., 2006). The experiment was performed in a light intensity (75 ± 1 lux) and temperature (21.6 ± 1.5 °C) controlled room, in which no-one except the participant was present during the task.

The participants were asked to remain seated during the whole experiment to minimize the effect of physical movement and changes in body posture on skin temperature regulation. To minimize the effect of prior physical exercise before arrival at the lab, it was ensured that the participants were seated at least 30 minutes prior to the start of the vigilance task. During this acclimatization period, temperature loggers were attached to the participant, as described below. Furthermore, multiple questionnaires were given, of which none were used for data analysis.

The participants seated in front of a computer and faced a thermographic camera placed on a tripod. Their face was mounted in a chinrest to guarantee minimal movement during the camera recordings. They were presented with a 10-minute vigilance task, described below. This task was rehearsed 1 minute by every participant before the start of the task. Prior and after the task, the participants rated their subjective sleepiness on the Stanford Sleepiness Scale (Hoddes, 1973).

Vigilance assessment

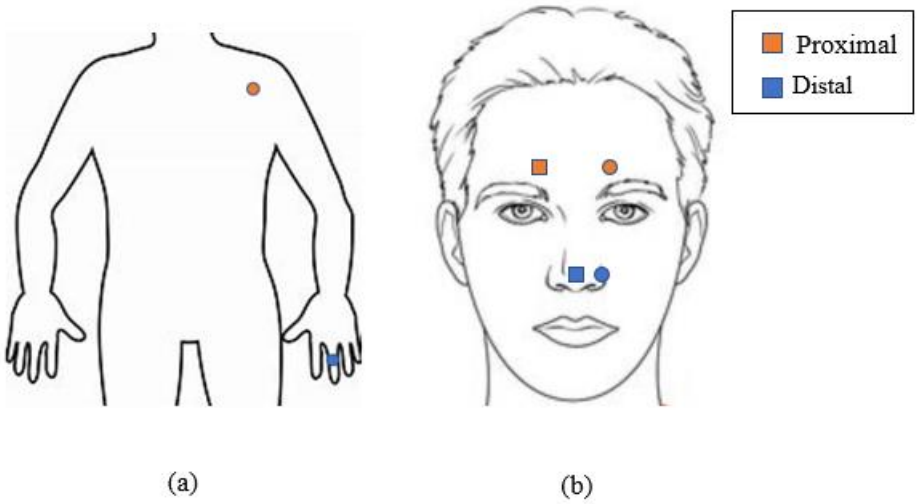
Vigilance was assessed using the 10-minute PVT. The PVT is a simple reaction time task that monitors vigilance, which is described in more detail in Dorrian et al. (2005). In brief, the participants were presented a black screen with a red box in the middle of the screen. A cue, the appearing of a millisecond (ms) counter inside the box, occurred at inter-stimulus interval (ISI), varying random between $1000 - 9000$ ms. Participants were instructed to press the spacebar of a keyboard with their dominant hand as fast as possible whenever the cue appeared. When pressed, the ms counter would stop and remain visible for 1 second, after which the next ISI commenced. Each PVT lasted 10 minutes and consisted of ~ 94 stimuli.

Temperature assessment

Skin temperature were 1). continuously measured from 4 regions of interest (ROI) on the skin by wireless temperature loggers and 2). recorded from the infrared spectrum using camera-based thermal imaging.

Figure 1

Skin temperature measurement location of the iButtons and Region of Interest (ROI) of the thermographic camera



Note. The skin temperature measurement locations on the body (a) and the face (b). The orange points are the proximal locations on the infraclavicular area below the clavicle (a) and the forehead above the eyebrow (b). The blue points are the distal locations on the dorsal side of the proximal phalange of the middle finger (a), the side of the nose (b) and the tip of the nose (b). The DPG is measured between the 2 closest distal-to-proximal locations, the finger-clavicle gradient and the nose-forehead gradient. The circles are the locations measured by the iButtons. The squares are the regions of interest recorded by the thermal camera.

The participants skin temperature was logged by 4 iButtons (type DS1922L, Maxim integrated, San Jose, USA), with samples of 30-sec intervals and a resolution of .0625 °C. This method has been validated and described in detail by van Marken Lichtenbelt et al. (2006). The iButtons were attached on two proximal positions, one placed on the infraclavicular area below the clavicle and one on the forehead above the eyebrow (Figure 1). iButtons were also attached on two distal locations, one on the dorsal side of the proximal phalange of the middle finger and one on the side of the nostril (Figure 1). They were attached on the non-dominant side of the body with tape (Fixomull stretch, BSN medical GmbH, Hamburg, Germany) to ensure minimal movement during the task. During the measurement, the clavicle was covered by a sweater to ensure equal conditions between participants.

Thermal images were recorded by a thermographic camera (FLIR E53 24°, *FLIR Systems Inc., Wilsonville, U.S.A.*) with an infrared resolution of 240 x 180 pixels, thermal sensitivity below 0.04 °C, and an accuracy of ± 2°C. The camera measured facial skin temperature with 0.033-sec sample interval and image frequency of 30 Hz.

Data analysis

All data was processed and analysed by using R-4.0.3.

Pre-processing. After the thermographic camera recordings, commercial software (Flirtools+, *FLIR Systems Inc., Wilsonville, U.S.A.*) was used to manually locate two ROIs on the face for data extraction, shown in figure 1. One proximal position located on the forehead above the eye pupil (~20 pixels upward) on the opposite side of the forehead of the iButton. One distal position located on the nose tip parallel to the located iButton on the nose (~10 pixels vertical shift). The tip of the nose was chosen instead of the side of the nose to minimize the effects from infrared radiation reflected by the cheek and avoiding the iButton sensor. Skin temperature values were defined by taking the mean temperature of each ROI (61 pixels area)

per frame. The pre-processed IRT data was split into temporal windows, each starting 2 seconds prior to the participants behavioural response per PVT stimuli.

Feature extraction. The DPGs were calculated over each stimuli or temporal window per participant. As mentioned, the DPG provides a reliable estimate of thermoregulatory changes and is therefore seen as the best predictor of vigilance levels (Kräuchi et al., 2000). A gradient could be calculated for a distal area minus a corresponding nearby proximal skin area (DPG) (Romeijn et al., 2012). For this study, three separate DPG were calculated. The first, iButton middle body DPG (DPG_iBmid) was calculated by the iButton between the measured DPG of the finger and the clavicle (Figure 1). This gradient has been validated by Raymann and van Someren (2007). The second, iButton face DPG (DPG_iBface), was calculated by the iButton between the measured DPG of the nose and the forehead (Figure 1). The third, IRT face DPG (DPG_IRT), was calculated by the infrared camera between the measured DPG of the nose and the forehead (Figure 1). The DPGs between the nose and forehead have not been validated in previous studies. However, the study of Bergersen (1993) has recognised the difference in the amount of AVA's between these locations. Bergersen found AVA's in the tip of the nose, suggesting a distal location, and no AVA's in the forehead, suggesting a proximal location. Furthermore, the face is often exposed, making it more easily to observe with IRT and is therefore used as a potential DPG.

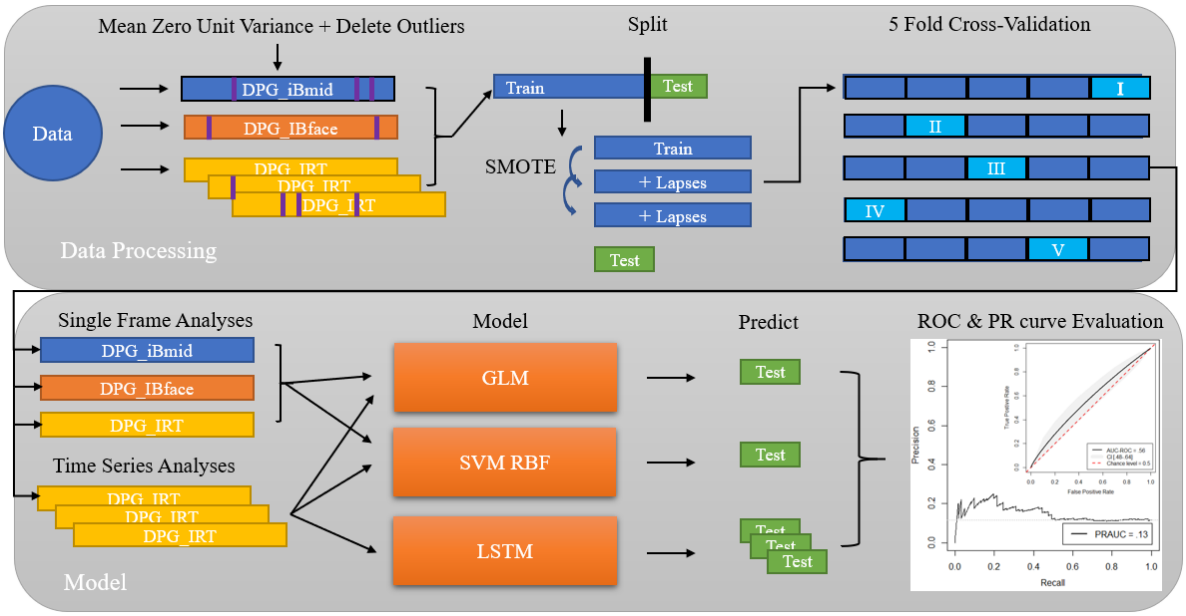
All DPG data was restricted to the behavioural response belonging to the stimuli presented during the trial. The DPG features corresponded to the last measured temperature readouts per stimuli response, or presented a sequence of 2-sec interval prior to this response.

Data processing. An overview of the data processing is shown in Figure 2. The DPG data was normalized to zero mean and unit variance within participant in order to avoid overfitting, account for normality and human error. The DPG_IRT data was manually extracted from the thermographic camera recordings, which could have resulted in potential human error.

Therefore, observations which exceeded ± 2 standard deviation from the zero mean were determined as outliers and deleted from the dataset. Furthermore the DPG_IRT data was checked for stationarity because of the assumption that time series are stationary.

Figure 2

An overview of the data processing, machine learning model and model evaluation.



Note. DPG_iBmid: iButton finger-clavicle gradient, DPG_iBface: iButton nose-forehead gradient. DPG_IRT: infrared camera nose-forehead gradient, GLM: Generalized Linear Model, SVM RBF: Support Vector Machine with Radial Basis Function, LSTM: Long Short-Term Memory model.

Per trial the level of vigilance was categorized in 2 states, either a lapse, loss in vigilance, or no-lapse, vigilant. Lapses were defined as omissions, reactions times (RTs) exceeding the 90th percentile of the distribution of the recorded RTs per participant (Dinges et al., 1997). In the dataset the number of no-lapses was far greater than the number of lapses,

causing a highly unbalanced dataset. This imbalance could impair the predictive ability of the algorithm because it pursues the overall classification accuracy (Liang et al., 2020). Meaning that it pays more attention to finding a no-lapse than a lapse. To overcome this problem, synthetic minority over-sampling technique (SMOTE) with 5 kernels was applied on the training set (Chawla et al., 2002). SMOTE is an over-sampling method in which the minority class, the lapses, is oversampled by creating artificial synthetic examples with k-nearest neighbours rather than normal replacement over-sampling (Ganganwar, 2012).

The data was split into a train and test set, where test data was withheld from training and otherwise. The train set contained the first 8 minutes of the PVT task per participant and the test set contained the last 2 minutes. In this way, the model could learn from the participants input. Five-fold cross validation was used to subsample validation data during training equivalent to the data evaluation of Zhang et al. (2016a).

Model

An overview of the three different types of models and their model evaluation is shown in Figure 2. The data analyses was split into Single Frame Analyses (SFA) and Time Series Analyses (TSA). The input of the SFA were the DPG_mid, DPG_iBface and DPG_IRT which correspond to the measured temperature readouts per stimuli response. The input of the TSA was the 2-second temporal window frame of the DPG_IRT. This input formed a time series sequence of 60 frames. In this study a time series was defined as a sequence of temperature data per trial per participant which did not overlap temperature data from other trials. Each PVT trial lasted 2 seconds or more depending on the ISI. No time series was used for the temperature data collected by the iButtons, because the 30-sec sampling interval could not be estimated inside the temporal window of interest.

Generalized Linear Model. A generalized linear model (GLM) was used to classify vigilance by SFA and TSA. A GLM was chosen because it is the standard data analysis in vigilance detection by DPG which was used in Romeijn et al., 2012. Therefore, this model was used as a baseline for vigilance classification and to differentiate between the iButtons and potential IRT estimation of vigilance in SFA. Although not commonly applied, some researchers have shown that with high temporal resolution data GLM could be used for TSA (Kristenen et al., 2017). The DPG used for TSA is measured by IRT (DPG_IRT) and considered to be of high temporal resolution. Therefore, the GLM model was also used for TSA vigilance classification.

Support Vector Machine with Radial Basis Function. A SVM with Radial Basis Function (RBF) was used to classify vigilance by SFA and TSA. The SVM applies a kernel function which extends the linear decision function to a non-linear high dimensional feature space function. A RBF kernels was chosen because it judges the similarity of two inputs by their Euclidian distance. It can detect anomalies in a window of sequences that are similar for other window frames. Therefore it can be promising for TSA (Rüping, 2001).

Long Short Time Memory. A LSTM neural network model was used to classify vigilance by TSA. In the research of Zhang et al. (2016a), LSTM models improved vigilance estimation compared to SVM RBF. Therefore, this study replicated the 3 stacked hidden layer LSTM model structure with tanh activation and sigmoid output by Zhang. Zhang designed two LSTM models who used EEG and EOG as input features. This research had one input feature and could therefore only adopt the model of Zhang which merged the EEG and EOG at feature level, the S-LSTM. A detailed description of the S-LSTM can be found in Zhang et al. (2016a). Different to Zhang et al. (2016a) was that no early stopping strategy was applied. Training stopped after 20 epochs and a batch-size of 10 sequences were given as input. Moreover, in

training, the RMSProp method with a fixed learning rate of 0.01 was used to optimize the binary cross entropy loss function.

Model evaluation. The Area Under The Receiver Operating Characteristics Curve (AUC-ROC) was used as a performance metrics during training of all models. The AUC-ROC is a performance measurement for classification problems and evaluates how good a model can distinguish between a lapse and no-lapse during training. The AUC-ROC was also used to evaluate models during testing. There was, however, a chance that the imbalanced test data could result in more optimistic and misleading AUC-ROC interpretation. Therefore the Precision-Recall Area Under Curve (PR-AUC) was also used for model evaluation. The Precision Recall (PR) curve has a strong relationship with the Receiver Operating Characteristics (ROC) curve (Davis & Goadrich, 2006) and focusses on the minority class, the lapses (Branco et al.,2015). It is therefore seen as an effective model evaluation for imbalanced binary classification models. In order to compare the models the Hanley and McNeil method was used to compare the AUC-ROC (Hanley & McNeil, 1982).

Results

Sample selection

The data contained 2719 samples of 29 participants. 45 outliers were deleted and the data per participant was split into training (2142 samples) and test-data (532 samples). Both lapse (*224 + 1100%,* 2464 samples) and no-lapse (*1918 +28.5%,* 2464 samples) data was increased by SMOTE oversampling in the training data.

Table 1

Single frame model performance of Generalized Linear Model and Support Vector Machine with Radial Basis Function on training and test data per temperature gradient

	GLM				SVM RBF			
	Train		Test		Train		Test	
	ROC (SD)	PR (SD)	ROC [CI]	PR	ROC (SD)	PR (SD)	ROC [CI]	PR
DPG_iBmid	*.62 (.02)*	*.64 (.01)*	***.58 [.51-.64]***	.13	***.75 (.02)***	***.72 (.02)***	.56 [.49-.64]	.10
DPG_iBface	.58 (.02)	.60 (.01)	.50 [.43-.58]	.11	.70 (.02)	.68 (.02)	.52 [.44-.60]	.12
DPG_IRT	.60 (.03)	.61 (.01)	.56 [.48-.63]	*.14*	.71 (.01)	*.68 (.02)*	.56 [.48-.64]	***.15***

Note. AUC-ROC (ROC) and PR-AUC (PR) per model and per temperature gradient (DPG). The models consist of the Generalized Linear Model (GLM) and Support Vector Machine with Radial Basis Function (SVM RBF). The DPG's consist of the iButtons (DPG_iBmid: finger-clavicle gradient and DPG_iBface: nose-forehead gradient) and the infrared camera (DPG_IRT: nose-forehead gradient). The GLM was as follows: $Lapse = DPG * time$, where lapse is a loss in vigilance, DPG is the temperature gradient in °C and time is the trial moment of PVT task in ms. SVM RBF classifies lapses by DPG and time with $C = 10$ and $\sigma = .5$.

Temperature measurement

Table 1 summarizes the model evaluation per DPG per SFA during training and testing. Both models used three types of DPG and time over task to classify lapses. For both models, the DPG between the finger and clavicle measured by the iButtons (DPG_iBmid) had an overall higher AUC-ROC compared to the DPG between the nose and forehead by the iButtons (DPG_iBface) and the infrared camera (DPG_IRT). The AUC-ROC suggested that the finger-

clavicle DPG was a better vigilance predictor than the nose-forehead gradient. On the other hand, the DPG_IRT had a higher PR-AUC than the DPG measured by the iButtons during testing.

Table 2

Difference between AUC-ROC of single frame GLM and SVM RBF per DPG

		GLM		SVM RBF		
		DPG_iBface	DPG_IRT	DPG_iBmid	DPG_iBface	DPG_IRT
		D *(p)*	D *(p)*	D *(p)*	D *(p)*	D *(p)*
GLM	DPG_iBmid	1.28 *(.20)*	.48 *(.63)*	.35 *(.72)*	.91 *(.36)*	.30 *(.76)*
	DPG_iBface		-.78 *(.44)*	-1.11 *(.27)*	-.79 *(.43)*	-.92 *(.36)*
	DPG_IRT			-.18 *(.86)*	.51 *(.61)*	-.05 *(.96)*
SVM RBF	DPG_iBmid				.69 *(.49)*	.10 *(.92)*
	DPG_iBface					-.68 *(.50)*

Note. Hanley and McNeil test for AUC-ROC with bootstrap = 2000. The difference is defined as $D = (AUCROC_1 - AUCROC_2)/s$, where s is the Standard Deviation of the bootstrap difference (Hanley & McNeil, 1982).

To determine whether there was a true difference per model performance per DPG, the Hanley and McNeil test was used to compare the AUC-ROC during testing (Hanley & McNeil, 1982). Table 2 shows the difference per model per DPG for SFA. No difference was found between using different DPG inputs for single frame model performances. These results

suggested that IRT classified vigilance as good as the validated finger-clavicle gradient measured by the iButtons.

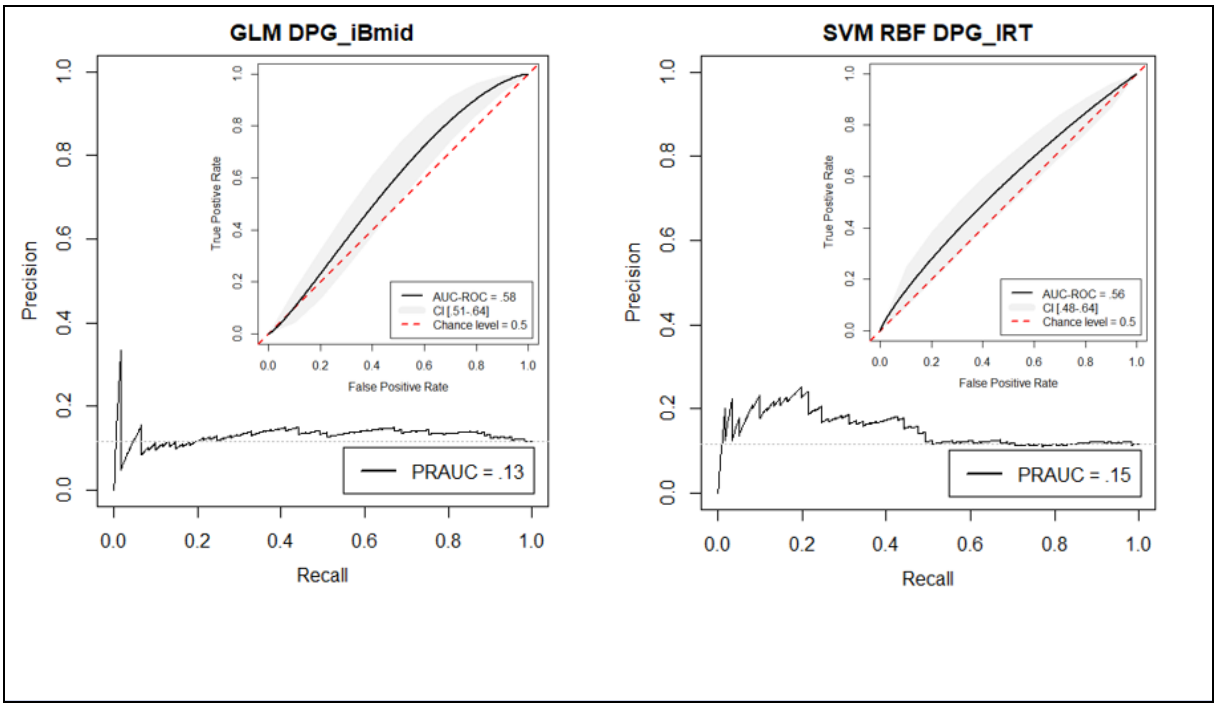
Single Frame Analysis

In order to find out whether machine learning can be applied for vigilance classification and even improve vigilance classification compared to standard linear classification, the standard linear analysis (GLM) was compared to non-linear machine learning analysis (SVM RBF) in SFA (Table 1). Overall, during the training the SVM RBF ($M = .72 SD = .03 AUC - ROC; M = .69 SD = .02$) outperformed the GLM ($M = .61 SD = .03 AUC - ROC; M = .62 SD = .02 PR - AUC$) on lapse detection, which suggested that the SVM RBF was an acceptable discriminator between lapses (*AUC-ROC > 0.7*) and that it improved vigilance classification over standard linear analysis (GLM). However, as mentioned both models did not significantly differ from each other based on AUC-ROC performance during testing (Table 2). The AUC-ROC of the GLM with infrared temperature input (DPG_IRT) showed almost no difference with the SVM RBF with the same input variables (DPG_IRT), $D = -0.05, p = .96$.

Thereby, during testing it was shown that both models were worse in classifying vigilance. In Figure 3 the best models of GLM and SVM RBF are shown, without taking into account the use of different DPG predictions. The GLM with the finger to clavicle gradient (DPG_iBmid) discriminated a lapse slightly above chance level of 0.5 (*AUC-ROC [CI] = .58 [.51-.64]*) and the PR-AUC was also around chance level of 0.09 (*PR-AUC = .13*). The same was seen for the infrared temperature of the nose to forehead gradient (DPG_IRT) for SVM RBF (*AUC-ROC[CI] = .56[.48-.64]; PR-AUC=.15*).

Figure 3

The PR curve and ROC curve of the Single Frame Analysis models with the highest AUC-ROC per model, GLM and SVM RBF



Note. Both plots show the Precision-Recall curve (black line) of a Single Frame Model, the dotted black line presents the PR curve change level. The inner plot shows Receiver Operating Characteristics curve (black line) of Single Frame Model, where the gray shape indicates the confidence interval of the AUC-ROC and the dotted red line presents the chance level. Left plot: SFA GLM DPG_iBmid; right plot: SFA SVM RBF DPG_IRT.

Time Series Analysis

Table 3 summarizes the model performance for TSA during training and testing. The GLM, SVM RBF and LSTM used a 2-second temporal window of the DPG_IRT as input data to classify lapses. This input met the time series assumption of stationarity ($DickeyFuller = -15.07, p = .01$) (Said & Dickey, 1984). Using DPG data over time (TSA) showed a higher

model performance during training for GLM and SVM RBF than SFA (Table 1). The SVM RBF showed promising results with a high AUC-ROC (*AUC-ROC (SD) = .90(.01)*) and PR-AUC (*PR-AUC (SD) = .84(.00)*), suggesting it was excellent at discrimination. The LSTM neural network model had the worst time series model performance (*ROC = .61 ; PR =.62)* during training, which was not as expected.

Table 3

Time series model performance of GLM, SVM RBF and LSTM. Bold marking the best scores per classifier per metric

	GLM		SVM RBF		LSTM	
	Train	Test	Train	Test	Train	Test
	ROC (SD)	ROC [CI]	ROC (SD)	ROC [CI]	ROC (SD)	ROC [CI]
	PR (SD)	*PR*	*PR (SD)*	*PR*	*PR (SD)*	*PR*
DPG_IRT	.72 (.02)	.54 [.46-.61]	**.90 (.01)**	**.56 [.48-.64]**	.61	.55 [.46-.63]
	.71 (.00)	***.12***	***.84 (.00)***	*.10*	*.62*	*.11*

Note. AUC-ROC (ROC) and PR-AUC (PR) scores per model; Generalized Linear Model (GLM), Support Vector Machine with Radial Basis Function (SVM RBF) and Long Short-Term Memory (LSTM) neural networks model. The GLM was as follows: $Lapse = DPG\_IRT_{t=t-0} + DPG\_IRT_{t=t-0.033} + DPG\_IRT_{t=t-0.066} + \cdots + DPG\_IRT_{t=t-2}$, where *lapse* was a loss in vigilance, *DPG_IRT* was the infrared camera nose to forehead temperature gradient (°C) at time *t*, the trial moment of PVT task. The SVM RBF and LSTM classified lapses by 2-second temporal window of DPG prior to the behavioural response per PVT data input. SVM RBF with $C = 10$ and $\sigma = .5$.

Figure 4 shows that the ROC curve and PR curve of the TSA models were around chance level, which matched their corresponding low AUC-ROC and PR-AUCs (Table 3). This was the same for the ROC curves and PR curves of the SFA (Figure 3), which was verified by the results in Table 4. No significant difference was found between the AUC-ROCs of the SFA and TSA, suggesting that time series did not improve vigilance classification.

Table 4

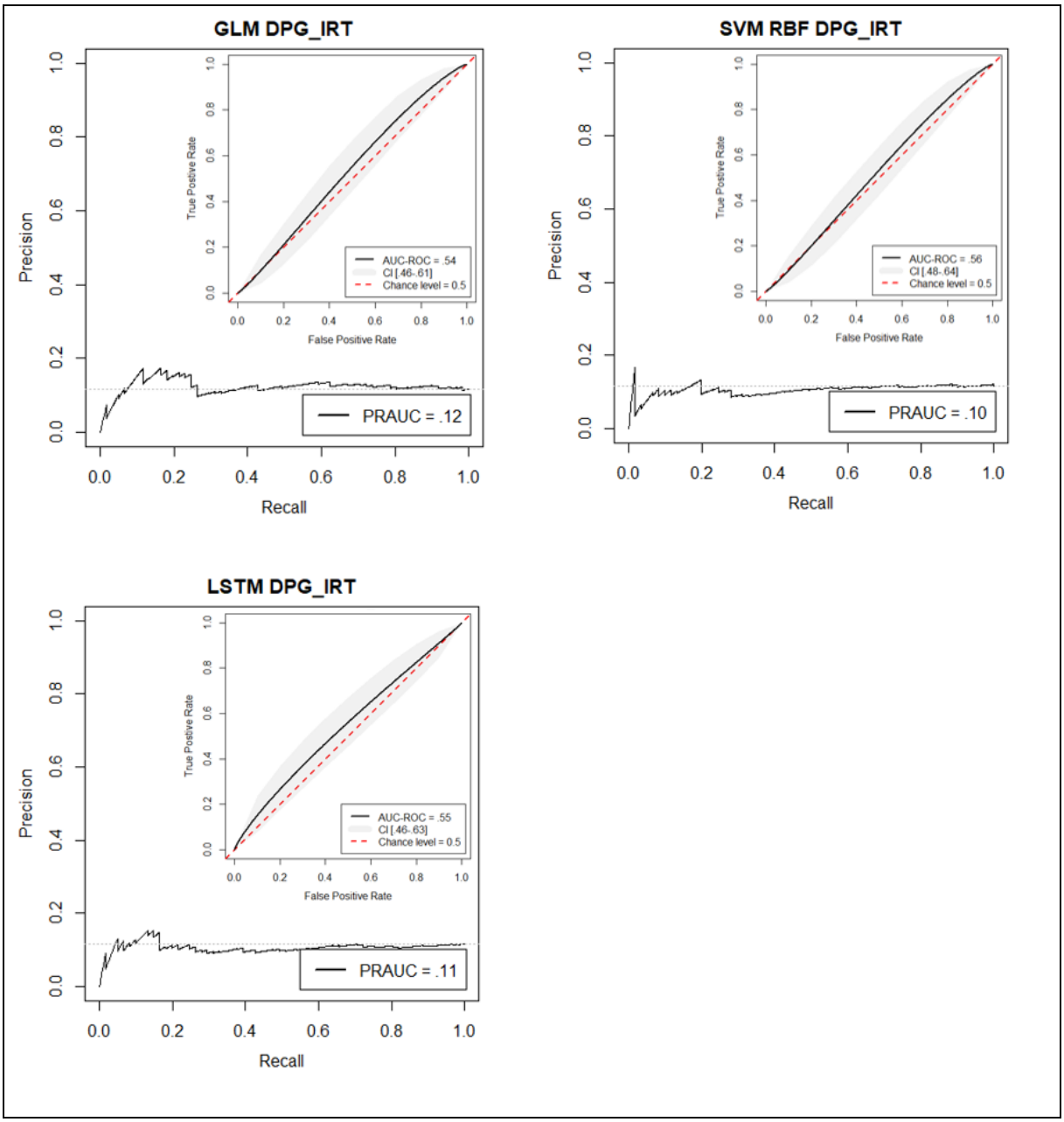
Difference between AUC-ROC per DPG_IRT model for Single Frame and Time Series Analysis

		Single frame analysis	Time Series Analysis		
		SVM RBF	GLM	SVM RBF	LSTM
		D *(p)*	D *(p)*	D *(p)*	D *(p)*
Single frame analysis:	GLM	-.05 *(.96)*	.32 *(.75)*	.46 *(.64)*	.16 *(.87)*
	SVM RBF		.40 *(.69)*	.52 *(.60)*	.20 *(.84)*
Time series analysis:	GLM			.18 *(.86)*	-.18 *(.86)*
	SVM RBF				-37 *(.71)*

Note. Hanley and Mcneil test for AUC-ROC with bootstrap = 2000. The difference is defined as $D = (AUCROC_1 - AUCROC_2)/s$, where s is the Standard Deviation of the bootstrap difference (Hanley & McNeil, 1982).

Figure 4

PR curve and ROC curve of the Time Series Analysis models (TSA).



Note. Both plots show the Precision-Recall (PR) curve (black line) of a TSA, the dotted black line presents the PR curve change level. The inner plot shows Receiver Operating Characteristics curve (black line) of TSA, where the gray shape indicates the confidence interval of the AUC-ROC and the dotted red line presents the chance level. Upper left plot: GLM; upper right plot: SVM RBF; lower left plot: LSTM.

Discussion

The aim of this study is to evaluate whether non-invasive continuous time-series measurements can detect and predict real-time vigilance levels.

The first objective is whether IRT can be used as a continuous non-invasive measurement to assess vigilance. The results found no performance difference between the models when iButtons or IRT was used as a predictor. These findings would have suggested that the DPG between the nose and forehead can be used as vigilance classifier and that IRT can be used as a continuous non-invasive measurement to assess vigilance. However, the models which were used to compare the vigilance measurements classified vigilance around chance level. The small differences between the performance of the models could be caused by models itself rather than the vigilance measurement.

Although not significant, the DPG_iBmid had a higher AUC-ROC and PR-AUC compared to the nose-forehead gradient. This could be caused by the difference in distal skin temperature. As mentioned, heat loss mainly takes place in distal skin locations because of the high density of AVA's (Charkoudian, 2003; Walløe, 2016). According to Walløe (2016), the AVAs in the hand, feet and the limbs are most responsible for heat loss regulation because these skin surfaces make up about 50% of the body surface. The nose contains only a small part of the skin surface and could therefore be too small to expose major temperature fluctuations. Nonetheless, the DPG_IRT had a higher AUC-ROC and PR-AUC compared to the DPG_iBface. It is possible that these minor temperature fluctuations in the nose were detected by the infrared camera because it had a higher temporal resolution and it was non-invasive. The iButtons were attached with tape which could affect the local heat transfer, and therefore the local thermal regulation (Quesada et al., 2015). This could suggest that IRT is even better in detecting temperature fluctuations than iButtons. Due to the lack of a control variable of the IRT measurement on the finger-clavicle gradient and the lack of a good baseline model, the results could not provide

conclusions about whether IRT can be used as a vigilance measurement. Nevertheless, it provides a new insight in the relationship between thermography and vigilance.

The second objective of this study is whether non-linear machine learning can improve vigilance classification. Results of the SFA showed that adding non-linearity (SVM RBF) did not improve classification over linear models (GLM), since no difference was found between models. Moreover, both models classified vigilance around chance level which was not as expected.

As mentioned, the GLM was used as a baseline analysis which was expected to be able to classify vigilance. However, during the training of the model performance of the GLM was low which resulted in a poor discrimination during testing (Hosmer et al., 2013). A probable cause for the low model performance during training was that the linear regression function was too simple. All features collected from the questionnaires were left out after five-fold cross-validation and only DPG and time were used as predictors to classify vigilance. A limitation to this research is that statistical features were not taken into account. These could have extended the formula leading to a better model performance.

The SVM RBF, on the other hand, showed on average an acceptable model performance (>0.7). Which suggested that it was able to detect patterns in DPG to classify vigilance. The model, however, did not generalize well on the test data and classified vigilance around chance level. Although cross-validation and SMOTE was applied there could be a possibility that the model still overfitted the data.

The third objective is whether TSA can improve vigilance classification. Results showed that model performance of GLM and SVM RBF improved when time series were added but did not generalize well on the test data. It is possible that adding time series was merely adding

more features than incorporating a TSA. Which could cause the models to overfit on the training data.

According to Rüping (2001) a linear model (GLM) cannot incorporate time dependency because this complexity can only be added by incorporating high level reasoning and background knowledge into the analysis. Whereas SVM can incorporate this complexity by using kernels. However, according to Zhang et al. (2016a), this is not enough to incorporate TSA. As mentioned earlier, vigilance is a dynamic process which should remember the history of the whole sequence, not only a window time frame. Only the LSTM could incorporate these sequences (Zhang et al., 2016a). Unfortunately, the LSTM had a low model performance during which resulted in poor vigilance classification.

Time is a very complex phenomena and very difficult to represent within the model. A reason for its complexity is that time can be presented in different representations (Rüping, 2001). In this study we chose to collect a time window of two seconds to incorporate within the model. Other forms of time series, such as different window sizes or statistical features could have improved the model and should be taken into account when considering further research.

At this moment no conclusions can be drawn from the results, since the baseline model with the validated DPG measurement was not able to classify vigilance. Limitations in data processing could have led to overall low model performance but also methodological factors, such as the young age of the participant group, could have played a role in lapse classification. A lapse was defined as a reaction time exceeding the 90th percentile of the distribution of RTs for PVT. Young people have faster reaction times than older people (Blatter et al., 2006), thus the lapse could be too biased to define a real loss in vigilance. A more distributed participant group could result in a more generalizable model representation.

Overall, this study was not able to classify vigilance by thermography. Machine learning models had not shown to improve data analysis and TSA had not improved vigilance classification. Nevertheless it showed that IRT could be a promising non-invasive continuous measurement to classify vigilance. This research is still at an early stage and future research should include a more distributed participant group and collect more statistical features. Ideally this research could provide new thoughts and insights in novel Time Series vigilance classification models. Whereby it can contribute in real-life situations such as monitoring vigilance in truck driving.

Acknowledgements

The author would like to thank Nico Romeijn for his supervision and Roderic Hillege for his shared expertise.

Literature

Aarts, L. T., Commandeur, J. J. F., Welsh, R., Niesen, S., Lerner, M., Thomas, P., Bos, N., & Davidse, R. J. (2016). Study on serious road traffic injuries 3 in the eu. *Publications Office of the European Union, Luxembourg.*

Abdelrahman, Y., Velloso, E., Dingler, T., Schmidt, A., & Vetere, F. (2017). Cognitive heat: exploring the usage of thermal imaging to unobtrusively estimate cognitive load. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):1–20.

Armanfard, N., Komeili, M., Reilly, J. P., & Pino, L. (2016). Vigilance lapse identification using sparse eeg electrode arrays. In *2016 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, pages 1–4. IEEE.

Aryal, A., Ghahramani, A., & Becerik-Gerber, B. (2017). Monitoring fatigue in construction workers using physiological measurements. *Automation in Construction*, 82:154-165

Bergasa, L. M., Nuevo, J., Sotelo, M. A., Barea, R., & Lopez, M. E. (2006). Real-time system for monitoring driver vigilance. *IEEE Transactions on Intelligent Transportation Systems*, 7(1):63–77.

Bergersen, T.K. (1993). A search for arteriovenous anastomoses in human skin using ultrasound Doppler. *Acta Physiologica*, *147*(2), p. 195-201.

Blatter, K., Graw, P., Münch, M., Knoblauch, V., Wirz-Justice, A., & Cajochen, C. (2006). Gender and age differences in psychomotor vigilance performance under differential sleep pressure conditions. *Behavioural brain research*, *168*(2), 312-317.

Charkoudian, N. (2003). Skin blood flow in adult human thermoregulation: how it works, when it does not, and why. In *Mayo clinic proceedings*, volume 78, pages 603–612. Elsevier.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357.

Davidson, P. R., Jones, R. D., & Peiris, M. T. (2007). Eeg-based lapse detection with high temporal resolution. *IEEE Transactions on Biomedical Engineering*, 54(5):832–839.

Davis, J., & Goadrich, M. (2006, June). The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning* (pp. 233-240).

Dinges, D. F. (1995). An overview of sleepiness and accidents. *Journal of sleep research*, 4:4-14.

Dinges, D. F., Pack, F., Williams, K., Gillen, K. A., Powell, J. W., Ott, G. E., ... & Pack, A. I. (1997). Cumulative sleepiness, mood disturbance, and psychomotor vigilance performance decrements during a week of sleep restricted to 4–5 hours per night. *Sleep*, *20*(4), 267-277.

Donald, F. M. (2008). The classification of vigilance tasks in the real world. *Ergonomics*, 51(11):1643–1655.

Dorrian, J., Rogers, N. L., & Dinges, D. F. (2005). *Psychomotor vigilance performance: Neurocognitive assay sensitive to sleep loss* (Doctoral dissertation, Marcel Dekker).

Fernández-Cuevas, I., Marins, J. C. B., Lastras, J. A., Carmona, P. M. G., Cano, S. P., García-Concepción, M. A., and Sillero-Quintana, M. (2015). Classification of factors influencing the use of infrared thermography in humans: A review. *Infrared Physics & Technology*, 71:28–55.

Fronczek, R., Raymann, R. J., Romeijn, N., Overeem, S., Fischer, M., Dijk, J. G. V., ... & Van Someren, E. J. (2008). Manipulation of core body and skin temperature improves vigilance and maintenance of wakefulness in narcolepsy. *Sleep*, *31*(2), 233-240.

Ganganwar, V. (2012). An overview of classification algorithms for imbalanced datasets. *International Journal of Emerging Technology and Advanced Engineering*, *2*(4), 42-47.

Geurts, P. (2001, September). Pattern extraction for time series classification. In *European Conference on Principles of Data Mining and Knowledge Discovery* (pp. 115-127). Springer, Berlin, Heidelberg.

Ghahramani, A., Castro, G., Becerik-Gerber, B., & Yu, X. (2016). Infrared thermography of human face for monitoring thermoregulation performance and estimating personal thermal comfort. *Building and Environment*, 109:1– 11.

Gilbert, D. G., Dibb, W. D., Plath, L. C., & Hiyane, S. G. (2000). Effects of nicotine and caffeine, separately and in combination, on EEG topography, mood, heart rate, cortisol, and vigilance. *Psychophysiology*, *37*(5), 583-595.

Gravetter, F. J., & Wallnau, L. B. (2013). Statistics for the behavioral sciences, London: Thomson Wadsworth.

Hanin, B. (2018). Which neural net architectures give rise to exploding and vanishing gradients? In *Advances in Neural Information Processing Systems*, pages 582-591

Hanley, J. A. & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143,* 29-36.

Hassoun, M. H. et al. (1995). *Fundamentals of artificial neural networks.* MITpress

Hasselberg, M. J., McMahon, J., & Parker, K. (2013). The validity, reliability, and utility of the ibutton R for measurement of body temperature circadian rhythms in sleep/wake research. *Sleep medicine*, 14(1):5–11.

Hoddes, E., Zarcone, V., Smythe, H., Phillips, R., & Dement, W.C. (1973). Quantification of sleepiness: a new approach. *Psychophysiology*, *10*(4), p. 431-436.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

Ioannou, S., Gallese, V., & Merla, A. (2014). Thermal infrared imaging in psychophysiology: potentialities and limits. *Psychophysiology*, 51(10):951– 963.

Kitney, R. I. (1975). An analysis of the nonlinear behaviour of the human thermal vasomotor control system. *Journal of Theoretical Biology*, *52*(1), 231-248.

Kräuchi, K., Cajochen, C., Werth, E., & Wirz-Justice, A. (1999). Warm feet promote the rapid onset of sleep. *Nature*, 401(6748):36–37.

Kräuchi, K., Cajochen, C., Werth, E., & Wirz-Justice, A. (2000). Functional link between distal vasodilation and sleep-onset latency? *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, 278(3):R741– R748.

Kräuchi, K. & Wirz-Justice, A. (1994). Circadian rhythm of heat production, heart rate, and skin and core temperature under unmasking conditions in men. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, 267(3):R819–R829.

Kristensen, E., Guerin-Dugué, A., & Rivet, B. (2017). Regularization and a general linear model for event-related potential estimation. *Behavior Research Methods*, *49*(6), 2255-2274.

Liang, X. W., Jiang, A. P., Li, T., Xue, Y. Y., & Wang, G. T. (2020). LR-SMOTE–An improved unbalanced data set oversampling based on K-means and SVM. *Knowledge-Based Systems*, 105845.

Lin, C.-T., Chuang, C.-H., Huang, C.-S., Tsai, S.-F., Lu, S.-W., Chen, Y.-H., & Ko, L.-W. (2014). Wireless and wearable eeg system for evaluating driver vigilance. *IEEE Transactions on biomedical circuits and systems*, 8(2):165– 176.

Marmarelis, V. Z. (1997). Modeling methology for nonlinear physiological systems. *Annals of biomedical engineering*, *25*(2), 239-251.

Masoudian, M. & Razavi, H. (2019). An investigation of the required vigilance for different occupations. *Safety Science*, 119:353–359.

Neapolitan, R. E., & Jiang, X. (2018). *Artificial intelligence: With an introduction to machine learning*. CRC Press.

Oken, B. S., Salinsky, M. C., & Elsas, S. (2006). Vigilance, alertness, or sustained attention: physiological basis and measurement. *Clinical neurophysiology*, 117(9):1885–1901.

Oudonesom, V. (2001). *Evaluation of techniques for vigilance measurements.* PhD thesis, Massachusetts Institute of Technology.

Parasuraman, R., Warm, J. S., & See, J. E. (1998). Brain systems of vigilance.

Pernet, C. R., Chauveau, N., Gaspar, C., & Rousselet, G. A. (2011). LIMO EEG: a toolbox for hierarchical LInear MOdeling of ElectroEncephaloGraphic data. *Computational intelligence and neuroscience, 2011*.

Puri, C., Olson, L., Pavlidis, I., Levine, J., & Starren, J. (2005). Stresscam: non-contact measurement of users' emotional states through thermal imaging. In *CHI'05 extended abstracts on Human factors in computing systems*, pages 1725–1728.

Quesada, J. I. P., Guillamón, N. M., de Anda, R. M. C. O., Psikuta, A., Annaheim, S., Rossi, R. M., ... & Palmer, R. S. (2015). Effect of perspiration on skin temperature measurements by infrared thermography and contact thermometry during aerobic cycling. *Infrared Physics & Technology*, *72*, 68-76.

Quinlan, P. T., Lane, J., Moore, K. L., Aspen, J., Rycroft, J. A., & O'Brien, D. C. (2000). The acute physiological and mood effects of tea and coffee: the role of caffeine level. *Pharmacology Biochemistry and Behavior*, *66*(1), 19-28.

Raymann, R. J., & Van Someren, E. J. (2007). Time-on-task impairment of psychomotor vigilance is affected by mild skin warming and changes with aging and insomnia. *Sleep*, *30*(1), 96-103.

Reyes, M. L., Lee, J. D., Liang, Y., Hoffman, J. D., & Huang, R. W. (2009). Capturing driver response to in-vehicle human-machine interface technologies using facial thermography.

Romeijn, N., Verweij, I. M., Koeleman, A., Mooij, A., Steimke, R., Virkkala, J., van der Werf, Y., & Van Someren, E. J. (2012). Cold hands, warm feet: sleep deprivation disrupts thermoregulation and its association with vigilance. *Sleep*, 35(12):1673–1683.

Rüping, S. (2001). *SVM kernels for time series analysis* (No. 2001, 43). Technical report.

Russell, S. J., & Norvig, P. (2016). Artificial intelligence: a modern approach. Malaysia.

Said, S. E., & Dickey, D. A. (1984). Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika*, *71*(3), 599-607.

Sano, A., Chen, W., Lopez-Martinez, D., Taylor, S., & Picard, R. W. (2018). Multimodal ambulatory sleep detection using lstm recurrent neural networks. *IEEE journal of biomedical and health informatics*, 23(4):1607–1617.

Schnupp, T., Heinze, C., Groß, H. M., & Golz, M. (2010). Long Short-Term Memory Training for the Assessment of Vigilance. *Biomed Tech*, *55*, 1.

Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Stemberger, J., Allison, R. S., & Schnell, T. (2010). Thermal imaging as a way to classify cognitive workload. In *2010 Canadian Conference on Computer and Robot Vision*, pages 231–238. IEEE.

Sundermeyer, M., Schlüter, R., & Ney, H. (2012). Lstm neural networks for language modeling. In *Thirteenth annual conference of the international speech communication association*.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Van Marken Lichtenbelt, W.D., Van Daanen, H.A., Wouters, L., Fronczek, R., Raymann, R.J., Severens, N.M., & Van Someren, E.J. (2006). Evaluation of wireless determination of skin temperature using iButtons. *Physiology & behavior*, *88*(4-5), p. 489-497.

Van Orden, K. F., Jung, T. P., & Makeig, S. (2000). Combined eye activity measures accurately estimate changes in sustained visual task performance. *Biological psychology*, *52*(3), 221-240.

Walløe, L. (2016). Arterio-venous anastomoses in the human skin and their role in temperature control. *Temperature*, 3(1):92–103.

Wei, Z.-P. & Lu, B.-L. (2012). Online vigilance analysis based on electrooculography. In *The 2012 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE.

Zhang, N., Zheng, W.-L., Liu, W., & Lu, B.-L. (2016a). Continuous vigilance estimation using lstm neural networks. In *International Conference on Neural Information Processing*, pages 530–537. Springer.

Zhang, Z., Luo, D., Rasim, Y., Li, Y., Meng, G., Xu, J., & Wang, C. (2016b). A vehicle active safety model: vehicle speed control based on driver vigilance detection using wearable eeg and sparse representation. *Sensors*, 16(2):242.

Zhao, W., Bhushan, A., Santamaria, A. D., Simon, M. G., & Davis, C. E. (2008). Machine learning: A crucial tool for sensor design. *Algorithms*, *1*(2), 130-152.

Zheng, W.-L. & Lu, B.-L. (2017). A multimodal approach to estimating vigilance using eeg and forehead eog. *Journal of neural engineering*, 14(2):026017.